U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
SYSTEMS DEVELOPMENT OFFICE
TECHNIQUES DEVELOPMENT LABORATORY

TDL OFFICE NOTE 78-6

COMPARATIVE VERIFICATION OF OPERATIONAL TWO TO SIX HOUR OBJECTIVE FORECASTS AND OFFICIAL NWS WATCHES OF SEVERE LOCAL STORMS

Jerome P. Charba and Stephen M. Burnham

April 1978

COMPARATIVE VERIFICATION OF OPERATIONAL TWO TO SIX HOUR OBJECTIVE FORECASTS AND OFFICIAL NWS WATCHES OF SEVERE LOCAL STORMS

Jerome P. Charba and Stephen M. Burnham

1. INTRODUCTION

Since 1974, 2-6 h probability forecasts of severe local storms (tornadoes, large hail, and damaging surface winds) developed by the Techniques Development Laboratory (TDL) have been transmitted from the National Meteorological Center (NMC) to the National Severe Storms Forecast Center (NSSFC) and other NWS forecasting offices. This product was developed with the primary aim of aiding NSSFC forecasters in issuing tornado and severe thunderstorm watches. The product is currently transmitted on the Request/Reply system under the heading FOUS80 during the spring and summer seasons. Four 2-6 h forecasts are issued daily, spanning the period 1700 to 0600 GMT (see Charba, 1977, for further details).

After four seasons of operational transmission of the forecasts, it should be beneficial to field forecasters to objectively evaluate their usefulness (or the potential thereof) as guidance. While both qualitative and quantitative verifications of the probabilities against reports of severe storm occurrences have been carried out in the past (Charba, 1975; 1977), the scores tell little about the value of the forecasts as guidance because we don't know how the official watches performed for the same period. This report discusses a quantitative comparative verification between TDL's objective forecasts and official NWS tornado and severe thunderstorm watches issued by NSSFC. The verification sample consists of all operational forecasts issued during the spring season (mid-March to mid-June) of 1977.

This study attempts to answer the following question: How do the objective forecasts compare with the watches in their ability to produce forecasts of practical value to the ultimate user, i.e., the general public? If they compare favorably with the watches in this respect, then it follows that the objective forecasts should have significant guidance value to a person issuing watches. Since the lead time and valid period of a forecast are related to its effectiveness or value, and since these time parameters differ between the two sets of forecasts, they are taken into account in the scoring.

We want to make it clear that the scores for the two sets of forecasts should not be used to judge the relative severe storm forecasting capability of the computer versus the person at the forecast desk. Such an interpretation of the scores would be unfair and improper because of the differences in the conditions or restrictions surrounding the two forecasting systems.

2. METHOD

A. Spatial Matching of the Two Forecast Sets

To properly conduct a comparative verification, one must be sure that the forecasts from each system are of the same type (e.g., probability or categorical) and that the event is the same for each of them. The 2-6 h objective forecasts are in the form of probability of occurrence of a storm event. The event for the 4-h period is an occurrence of one or more tornadoes, hail $\geq$ 3/4 in ($\sim$2 cm) diameter, surface wind gusts of 50 kt ($\sim$93 km/h) or wind damage in square areas of about 85 n mi ($\sim$160 km) on a side during a 4-h period (Fig. 1). An official watch is in the form of a statement that one or more of these severe weather types are likely to occur within a delineated area, generally rectangular in shape (Fig. 1). Watches are issued separately for thunderstorms expected to produce tornadoes and for thunderstorms expected to produce only large hail or damaging surface wind gusts. The valid period of a watch is generally around 5-6 hours and the area covered averages about 22,000 n mi^2 ($\sim$75,000 km^2).

Thus, it is seen that the two sets of forecasts have differences between them in both the form of the forecast and in the nature of the event. The difference in form was remedied by converting the probabilities into categorical YES/NO statements. The valid area corresponding to the YES/NO forecast is in the shape of a square 40-45 n mi ($\sim$75-85 km) on a side. (Recall that the valid area for the analogous probability forecast also forms a square but it covers 4 times as much area.) The watch rectangles were also broken into these small squares and the event was defined as the occurrence or nonoccurrence of one or more tornadoes, hail, or damaging wind within these squares (see Fig. 1). Individual squares were defined to lie within the watch if their center points fell within the watch area. Later, it is shown that this method of approximating the total watch areas had a negligible effect on their overall verification scores.

In the watch verification, both the tornado and severe thunderstorm watches were used and no distinction was made between YES forecasts based upon each type. Thus, a tornado, hail, or severe wind occurrence verifies a YES forecast regardless whether the YES is based upon a tornado or a severe thunderstorm watch. In this way, the verification of the watches is consistent with that of the objective forecasts.

The probabilities were converted into the "best" obtainable categorical forecasts by empirical adjustment of a threshold probability value. When the probability is greater than or equal to this threshold, the categorical forecast is for an occurrence of the event, i.e., a YES forecast; otherwise the forecast is for a nonoccurrence (NO forecast). A "first guess" at the "best" threshold value was obtained by trial and error using a sample of forecasts obtained by applying the operational regression equations to the developmental data sample (spring seasons of 1974-76).

This threshold is the one which yields the best combination of values of Threat score[1] (Palmer and Allen, 1949) and bias.[2]

The ultimate choices of the threshold values were made after we had seen both the initial verification results based upon the first guess choices and the results of the watch verifications. Final adjustments to the thresholds were made after examining the verification results of the watches in order to arrive at as much consistency as possible between the two systems in the number of storm occurrences correctly forecast (hits) relative to the number of incorrect YES forecasts (false alarms). An experiment aimed at testing the validity of this procedure for converting from the probability into categorical forecasts is discussed in the Appendix.

The total area over which the forecasts were verified is shown in Fig. 1. Since different regression equations are used to produce the forecasts in the Gulf and non-Gulf regions and also because of the known difference in forecasting difficulty across these regions (Galway, 1975), the verification statistics will be presented separately for each area. For reasons explained in the next subsection, only the second, third, and fourth of the daily objective forecasts, i.e., those issued at 1845, 2145, and 0045 GMT as illustrated in Fig. 2, were used in this comparative verification. The respective threshold probability values used were 23, 22, and 15% for the Gulf region and 22, 22, and 18% for the non-Gulf region.

B. Temporal Matching of the Two Forecast Sets

The objective forecasts are issued at 3-hourly intervals (Fig. 2) while watches are issued at any time. In addition, the valid time periods of watches generally vary from one watch to another. Since the issue times and valid periods of the two sets of forecasts do not match, some procedure had to be devised to overcome these differences. After examining the watch data, it soon became apparent that it would be futile to attempt even a rough matching of their time parameters with those of the objective forecasts. Consequently, we rejected the idea of attempting to match individual issue times and valid periods and, instead, verified all forecasts falling within a fixed period of the day. Noting that very few of the watches had valid periods falling much before 2000 GMT and after 0600 GMT, we selected the period 2000-0600 GMT as the verification period. Another factor prompting this particular period for verification was that 2000 GMT marks the beginning of the valid period of the second objective forecast and 0600 GMT ends the valid period of the last objective forecast of the day (see Fig. 2). Thus, all objective forecasts and watches whose valid periods fell totally within the verification period were verified together.

---

[1] The Threat score, also called critical success index (CSI) by Donaldson (1975), is the number of correct YES forecasts divided by the sum of the correct YES, incorrect YES, and incorrect NO forecasts.

[2] The bias is defined as the number of YES forecasts divided by the number of occurrences of the event.

A problem that remains, however, is the first objective forecast of the day and all watches whose valid periods fall only partly within the verification period (Fig. 2). The final decision was to eliminate any forecast in the two sets whose valid period overlapped the verification period by one hour or less. This overlap criterion was arrived at on the basis of two considerations. First, we did not want to include the first objective forecast in the comparative verification because very few watches in the sample spanned its valid period. This one hour overlap criterion eliminates it. The second criterion was that the average valid time (i.e., beginning of the valid period) of the earliest watches to be included in the verification was to be near 2000 GMT and the average ending time of the latest watches was to be near 0600 GMT. The values resulting from the one hour criterion were 1950 and 0423 GMT, respectively. Frequency histograms which show the actual distributions of the valid and ending times are shown in Figs. 3a and 3b, respectively.

Summarizing, the verification period for the objective forecasts was fixed at 2000-0600 GMT while that for the watches varied from one day to the next. It is important to note that this variable verification period in the case of the watches applied only for the case in which YES forecasts are being verified. Conversely, when we considered the number of individual storm occurrences correctly forecast (hits) only those occurrences falling within period 2000-0600 GMT were involved. This means that exactly the same set of storm reports were involved in the determination of the number of hits scored by each system. In the case of the YES forecasts, the sets of individual storm occurrences used in their verification were slightly different for the two sets of forecasts because of the variable verification period in the case of the watches.

The period of the 1977 season involved in the verification ran from March 16 to June 15. The sample included 85 days after removal of days on which one or more of the objective forecasts were not available. For this 85 day period, 131 watches were issued on 48 days giving an average of 2.7 watches per day. The number of issuances of the objective forecasts was 85 x 3 or 255 but, of course, on some days all the individual forecasts fell in the NO category.

C. Verification Scores

Forecast Timing Not Considered

When dealing with YES/NO categorical forecasts and events, the standard procedure is to use scores that can be computed from 2 x 2 contingency tables. Such tables can readily be developed for the objective forecasts but not for the watches. The problem with the watches stems from the fact that sometimes no watches are in effect while at other times several are in effect with each having a different issue time. An attempt to develop a 2 x 2 contingency table is met with the obstacle of not knowing how to tabulate the storm occurrences that fall spatially outside all the watches in effect. The problem is compounded for the case in which the missed storm occurs at a time when two or more watches are in effect. In other

words, observed events can be defined only within the temporal and spatial confines of individual watches; outside the watches events cannot be defined because a valid period is not prescribed.

While some of the common verification scores could not be used, the performance of the sets of forecasts could be measured in terms of two separate quantities. One, called the probability of detection (POD) by Donaldson (1975), is defined as

$$POD = \frac{x}{x + y}, \tag{1}$$

where x is the total number of hits (i.e., the number of reported storm occurrences temporally and spatially captured by YES forecasts) and y is the number of occurrences falling outside the YES forecasts (misses). The other quantity, called the false alarm ratio (FAR) by Donaldson (1975), is defined as

$$FAR = \frac{z'}{x' + z'}, \tag{2}$$

where z' is the number of incorrect YES forecasts and x' is the number of correct YES forecasts. Primed symbols are used in (2) to distinguish correct or incorrect YES forecasts from hits or misses used in the POD.[1] It's worth noting that one correct YES forecast could result in two or more hits. The reason is that a YES forecast may be verified by several reported storm occurrences each of which is a hit. Also noteworthy is that 1-FAR is the percent correct in the case of the YES forecasts.

Forecast Timing Incorporated

According to the Weather Service Operations Manual (National Weather Service, 1977), the purpose of tornado and severe thunderstorm watches is to inform local NWS offices to be prepared to issue a warning, to activate storm spotter groups and civil defense offices, and to advise the public to be prepared to act in the event a warning is issued (also see Mogil and Groper, 1977). Obviously, time is required, especially for spotters to take their posts and to inform the public. Therefore, it would be highly desirable for each watch issued to have an adequate lead time (LT), the time between the beginning of the valid period and the time of issue. If one accepts the premise that the longer the time available for dissemination the greater the likelihood that users would become aware of the watch, the longer the LT the better. (In practice, a forecaster must make a trade-off between lead time and forecast accuracy.) Additionally, an accurate forecast should be more effective the shorter its valid period (VP), again assuming an acceptable level

---

[1]The POD would be identical to the prefigurance if observed predictand events were being considered instead of individual storm occurrences while 1-FAR is identical to the post agreement (Panofsky and Brier, 1968).

of accuracy is retained. Therefore, it seems appropriate to weigh hits and correct YES forecasts according to these time parameters. In accord with these considerations, one requirement for a weighting function is that it give credit in direct proportion to LT and in inverse proportion to VP.

Unfortunately, the current state of forecasting severe storms is not advanced to the point that all watches issued have an adequate LT. In fact, the situation is not uncommon that a watch is not issued until after tornadoes or severe thunderstorms have been reported. Since the weighting criterion stated above gives no credit to hits or correct YES forecasts in this case, an additional criterion must be provided for in the weighting function. It seems logical that the credit given be in proportion to the difference between the issue time of the watch and occurrence time of the storm. This time difference, called the projection time (PT), implies a lead time of sorts. The PT must be divided by the VP in the weight function so that excessive and undue credit is not given in the case of a very long PT associated with a watch with an even longer VP.

For the general case of a watch with a nonzero LT and PT, the weight function, w, was defined simply as

$$w = \alpha \frac{LT}{VP} + \gamma \frac{PT}{VP} \quad , \tag{3}$$

where the empirical factors α and γ allow for adjustment of the relative weighting of the two terms. Regarding the relative values that should be assigned the parameters α and γ, it is clear that the first term in (3) should be given greater weight than the second. For example, a nonzero LT is essential for those who must take protective action, for spotters, and for dissemination of the watch to the general public. The PT, on the other hand, would be worth much less to users because the time available is not known until after the storm has occurred![1]

Since a determination of the optimum relative weighting of the two terms in (3) goes beyond the scope of this study, we arbitrarily chose the values of 2.0 for α and 1.0 for γ. For these choices, the values taken on by w for different combinations of LT, PT, and VP are shown in Table 1. Note that in the case of a watch with a 2-h LT and 5-h VP, which we arbitrarily call an "ideal watch," the w values may be greater than one. On the other hand, watches with zero lead time result in w values in the range zero to one. It's interesting that w values corresponding to TDL's objective forecasts are almost as high as those for the "ideal watch." This result comes about because the objective forecasts and the ideal watch exhibit similar values in the ratio LT/VP. Thus, this ratio dominates the weighting function as it properly should according to the preceding

---

[1]For the general case in which LT and PT are both nonzero, it seems that the contribution from the second term in (3) should be dependent upon the value of LT. It could also be argued that the maximum value allowed for PT should be less than VP and that this maximum be dependent upon LT.

arguments. Therefore, the values 2.0 and 1.0 for α and γ, respectively, result in reasonable values for w and were used in this study. (It's worth adding, though, that the exact relative values assigned to α and γ should not be of great concern here because the weighting is applied equally to the objective forecasts and watches.)

Recalling that a hit refers to an individual storm occurrence correctly forecast, the total number of weighted hits, x_w, is

$$x_w = \sum_{i=1}^{N} w_i \quad , \tag{4}$$

where N is the total number of hits. Upon substituting (3) into (4), we obtain

$$x_w = \sum_{i=1}^{N} \frac{2LT_i + PT_i}{VP_i} \quad .$$

Substituting x_w for x in (1) gives the weighted POD written as

$$POD_w = \frac{\sum_{i=1}^{N} \frac{2LT_i + PT_i}{VP_i}}{\sum_{i=1}^{N} \frac{2LT_i + PT_i}{VP_i} + y} \quad . \tag{5}$$

Similarly, the weighted FAR is expressed as

$$FAR_w = \frac{\sum_{j=1}^{M} \frac{2LT_j + PT_j}{VP_j}}{\sum_{j=1}^{M} \frac{2LT_j + PT_j}{VP_j} + z'} \tag{6}$$

where M is the total number of correct YES forecasts.

Recalling that a YES forecast may be verified by more than one storm occurrence, one must choose a specific occurrence to evaluate PT in (6). Since the utility of these forecasts is directly related to the advance warning, we chose the first or earliest occurrence to evaluate this implied lead time variable.

3. RESULTS AND DISCUSSION

The objective forecasts and watches were verified against reports of severe storm occurrences gathered and carefully checked by NSSFC. Table 2a shows the unweighted number of hits and POD scored by each set of forecasts in the Gulf and non-Gulf regions. Note that the POD values for the watches are substantially larger than those for the objective forecasts in each region. The POD values for the case in which the watch areas were not approximated by the small verification squares (see Fig. 1) are shown in parentheses. Clearly, the approximation of the watch areas by these squares has a negligible overall effect on the POD.

Table 2b shows the performance of the YES forecasts for the two sets. The percent correct and FAR values show that the watches, again, outperformed the objective YES forecasts but the margin is not as large as it was for the POD. It is interesting to note that in the non-Gulf region the total number of YES forecasts was about the same for the two sets; in the Gulf region, on the other hand, the watches produced almost twice as many YES forecasts. Also, the objective forecasts exhibit a greater degradation in precent correct and FAR as one proceeds from the non-Gulf to the Gulf region.

Although the results just discussed show that the watches netted more hits and correct YES forecasts, we must consider the time of occurrence of the hits relative to the forecast issue time, the LT (lead time), and the VP (valid period). Figs. 4a and 4b show frequency distributions of hits as a function of time for the objective forecasts and watches, respectively, for the non-Gulf region. (Corresponding figures are not shown for the Gulf region because the samples are too small to be meaningful.) Analogous frequency histograms for correct YES forecasts are shown in Figs. 5a and 5b. In all of these figures, zero along the time axis corresponds to the forecast issue time. Plotted above each of the histograms is a schematic depicting the mean LT, PT, and VP, i.e., $\overline{LT}$, $\overline{PT}$, and $\overline{VP}$, respectively.

Several features in Figs. 4a and 4b are worth noting. One is that the distribution of hits for the objective forecasts is rather uniform over the 4-h VP (Fig. 4a) with 51% of the hits occurring during the first half. In the case of the watches (Fig. 4b), a clear majority of hits (67%) occurred during the first half of the VP. From a different perspective, this result is illustrated by the position of $\overline{PT}$ relative to $\overline{VP}_c$ (or $\overline{VP}c$), the center of the $\overline{VP}$. Note further that very few storms (1%) were correctly forecast by watches extending beyond 6 1/2 h into the future from issue time.

Another significant feature illustrated in Figs. 4a and 4b is that the ratio LT/VP is much higher for the objective forecasts than it is for the watches. It has a value of 0.31 for the objective forecasts and 0.11 for the watches in this non-Gulf region. In addition, while the PT value for the watches is larger than it is for the objective forecasts, the sum, LT + PT, remains larger for the objective forecasts. The differences in these time variables will be reflected in the weights applied to hits and correct YES forecasts.

The histograms for the YES forecasts (Figs. 5a and 5b) are quite similar to those for the hits in all respects except one. The difference is that the frequency distribution is shifted slightly toward smaller projections from forecast issue time. This shift reflects the fact that the first storm verifying a YES forecast was used for evaluating the PT. Thus, the PT corresponding to the YES forecasts in both sets is somewhat smaller than it was for hits.

Statistics on the weighted hits and correct YES forecasts are given in Tables 3a and 3b, respectively. A comparison of Tables 2a and 3a reveals that the number of weighted hits, and therefore the POD, increased a small amount over the unweighted hits in the case of the objective forecasts. For the watches, on the other hand, the total number of weighted hits dropped substantially from the unweighted total. In fact, the number of weighted hits in the Gulf and non-Gulf regions combined for the objective forecasts (211.6) now exceeds the combined total (208.5) scored by the watches. The impact the individual weights had on the total number of weighted hits is indicated by the average weight values given in the right hand column of Table 3a. The average values for the objective forecasts are seen to be almost twice as large as those for the watches.

The impact of weighting on the correct YES forecasts is similar to that just discussed for hits (Table 3b). For these weighted YES's, the objective forecasts now yield a better percent correct and FAR than do the watches in the non-Gulf region; in the Gulf region, the watches continue to have an edge. When the two regions are combined, the overall scores are about the same for the two systems.

The basic question we had hoped to answer in this study is, what do the results of this comparative verification say about the value of the objective forecasts as guidance to field forecasters? The unweighted scores cannot be used to base an answer to this question because they do not account for significant differences in lead times and valid periods between the two sets of forecasts. The weighted scores, which take these differences into account in a manner consistent with the officially stated purpose of watches, should provide a much better representation of the forecasts' relative warning effectiveness as it relates to the time parameters of concern. The weighted scores indicate that the performance of the objective forecasts was comparable to that of the watches. Thus, we are led to the conclusion that the objective forecasts should have significant guidance value to the watch forecasters.

As a final note, it may be worth recalling and expanding upon a point made in the introduction to this paper. It is that the verification scores just discussed should not and, indeed, cannot be used to judge the relative severe storm forecasting capability of TDL's objective/statistical method versus that of the forecasters at NSSFC. A very basic reason for this is that the conditions or restrictions imposed upon each system are greatly different. Among these conditions are: (1) the objective forecasts are available to NSSFC forecasters as guidance when they issue watches; (2) these forecasters have almost complete flexibility in deciding upon issue times, lead times, and valid periods for their watches while these time parameters are fixed in the case of the objective forecasts; (3) the field forecasters use much more predictive information than is currently built into our objective forecasting model--such information as can be gleaned from radar reports, satellite data, real time storm reports, etc.

4. SUMMARY AND CONCLUSIONS

This comparative verification was conducted with the aim of assessing the value of the objective forecasts as guidance for the issuance of watches. The assumption was that if the verification scores of the objective forecasts compare favorably with those of the watches then the field forecasters should be able to make effective use of these forecasts as guidance.

Since the nature of the objective forecasts versus that of the watches is different in many ways, we had to design separate verification schemes for each set of forecasts. In the process of developing these schemes many obstacles standing in the way of one-to-one comparisons between the two sets of forecasts were encountered. While certain assumptions or approximations had to be applied in order to circumvent these obstacles they had to be relatively minor so as to not jeopardize the propriety of the comparative verification.

A major difference between the two sets of forecasts, whose accommodation required considerable effort, involves forecast timing parameters, i.e., the lead times, valid periods, and "projection times." Differences in these time parameters were accounted for by incorporating these parameters into the verification scores. This was done by weighting the individual hits and correct YES forecasts according to a simple function wherein these time parameters were the independent variables. The function weighs the successful forecasts in direct proportion to the lead time and, to a lesser extent, the projection time and in inverse proportion to the valid period. While the precise relative importance assigned the individual terms in the function was somewhat arbitrary, the choices of relative weights were consistent with the officially stated purpose of watches. Of course, the whole question of watch effectiveness or usefulness is a major one and it deserves considerable attention all to itself.

The verification scores, namely the POD and FAR, were better for the watches than they were for the objective forecasts when hits and correct YES forecasts were not weighted. When weighting was applied, these scores were very nearly the same for the objective forecasts as they were for the watches. The weighting favored the objective forecasts mainly because the ratio of their lead time to valid period was much higher than it was for the watches--almost three times as large.

Several additional findings emerged as by-products of the verifications. One is that although the watches had an average lead time of 35 min, the range of their lead times was rather large. For instance, 28 of the 131 watches in the sample (21%) had zero lead time while four (3%) had lead times of 2 h or more. Another finding is that a rather large majority of hits scored by the watches (two-thirds of them) occurred between the issue time and the mid-point of the average valid period. Also, only 1% of all hits scored by the watches occurred beyond 6 1/2 h after issue time, even though the average watch valid period projected just beyond 6 h. In the case of the objective forecasts, the distribution of hits was very uniform over the 4-h valid period with about 51% falling in the first half.

Since the weighted scores of the objective forecasts were about the same as those of the watches, our conclusion is that the objective forecasts should have considerable guidance value to the NSSFC forecaster. However, the amount of effort we had to expend in order to make the verifications comparable also suggest that a forecaster could have some difficulty in using the objective forecasts as guidance. One of the main incompatibilities between the two sets of forecasts is that the watches project farther into the future due mainly to their longer valid periods. The results show, however, that the number of hits scored near the far end of watches with long projections is negligible. Therefore, if the projections of watches were restricted to, say, 6 1/2 h there would be virtually no loss in the number of hits and the objective forecasts could become more useful as guidance. Another concern is that, since NSSFC forecasters must be alert to the possible issuance of a needed watch at any instant of time, the three-hourly issuance of the objective guidance forecasts is probably insufficient. A shortening of the time gap between the guidance to a one hour interval should satisfy the requirement. Indeed, such a schedule is operationally possible and it merits serious consideration.

5. ACKNOWLEDGMENTS

We would like to thank the following individuals for their contributions to this work. Harry Swenson and Horace Hudson of NSSFC provided the severe storm and watch data, Denis Sakelaris aided in the preparation of the illustrations, William Griner performed some of the computations, and Barbara Howerton did the typing.

6. REFERENCES

Charba, J. P., 1975: Operational scheme for short range forecasts of severe local weather. Preprints, Ninth Conf. on Severe Local Storms, Amer. Meteor. Soc., Boston, Mass., 226-231.

_____, 1977: Operational system for predicting severe local storms two to six hours in advance. NOAA Tech. Memo., NWS TDL-65, 36 pp.

Donaldson, R. J., Jr., R. M. Dyer, and M. J. Kraus, 1975: An objective evaluator of techniques for predicting severe weather events. Preprints, Ninth Conf. on Severe Local Storms, Amer. Meteor. Soc., Boston, Mass., 321-326.

Galway, J. G., 1975: Relationship of tornado deaths to severe weather watch areas. Mon. Wea. Rev., 103, 737-741.

Mogil, H. M., and H. S. Groper, 1977: NWS's severe local storm warning and disaster preparedness programs. Bull. Amer. Meteor. Soc., 58, 318-329.

National Weather Service, 1977: Severe local storm warning. Weather Service Operations Manual, National Weather Service, Silver Spring, Md., Ch. C-40.

Palmer, W. C., and R. A. Allen, 1949: Note on the accuracy of forecasts concerning the rain problem. U. S. Weather Bureau (unpublished notes).

Panofsky, H.A., and G. W. Brier, 1968: Some Applications of Statistics to Meteorology. Penn. State University, University Park, Penn., 224 pp.

APPENDIX

EXPERIMENT TESTING CONVERSION FROM PROBABILITY TO CATEGORICAL FORECASTS

It may be useful to call attention to an experiment designed to test the propriety of our method of converting the severe storm probabilities into categorical forecasts. A new, experimental regression equation was derived using predictand areas of the identical size and shape used in this verification. (Recall that the operational probabilities were valid for squares with areas four times as large as those used in the verification. The developmental sample was the same one used to develop the operational equations. Probability forecasts based on this experimental equation were developed from the dependent sample as well as from the independent sample used in this verification. These probabilities were then converted into categorical forecasts following the procedure used in this study and verified within areas of the same size used in the development.

We were surprised to find that the Threat score yielded by forecasts from this experimental equation was higher than that exhibited by forecasts using the analogous operational equation. The improvement in Threat score was 11% in the case of the dependent sample and 17% for the independent sample. While these differences in Threat score are not large, they cannot be disregarded. Unfortunately, a likely cause for their appearance has not been found.

Table 1. Weight values (w) for given values of lead time (LT), projection time (PT), and valid period (VP).

	LT (h)	PT (h)	VP (h)	w
Ideal Watch	2	4½	5	1.3
	2	2	5	0.8
	2	7	5	1.8
Typical Watch	½	2 3/4	5½	0.7
	½	0	5½	0.2
	½	5½	5½	1.2
Watch with Zero LT	0	2½	5	0.5
	0	0	5	0.0
	0	5	5	1.0
Objective Forecasts	1¼	2	4	1.1
	1¼	0	4	0.6
	1¼	4	4	1.6

Table 2a. Number of hits and probability of detection (POD) scored by the objective forecasts and watches on all storms occurring during the period 2000-0600 GMT. The POD values in parentheses for the watches pertain to the case in which the watch areas are not approximated by the small verification squares.

Forecasts	Region	No. of Occurrences	No. of Hits	POD
Objective	Gulf	135	13	0.10
Objective	Non-Gulf	701	179	0.26
Watches	Gulf	135	44	0.33 (0.33)
Watches	Non-Gulf	701	297	0.42 (0.43)

Table 2b. Results of verification of YES forecasts valid partly or totally within the period 2000-0600 GMT.

Forecasts	Region	No. YES Fcsts	No. Correct	Percent Correct	FAR
Objective	Gulf	186	8	4.3	0.96
Objective	Non-Gulf	1484	135	9.1	0.91
Watches	Gulf	369	39	10.6	0.89
Watches	Non-Gulf	1511	202	13.4	0.87

Table 3a. Same as Table 2a except that the hits are weighted.

Forecasts	Region	No. Occurrences	No. Hits	POD	Avg. Weight
Objective	Gulf	135	15.3	0.11	1.18
Objective	Non-Gulf	701	196.3	0.28	1.10
Watches	Gulf	135	27.3	0.20	0.62
Watches	Non-Gulf	701	181.2	0.26	0.61

Table 3b. Same as Table 2b except that correct YES forecasts are weighted.

Forecasts	Region	No. Yes Fcsts.	No. Correct	Percent Correct	FAR	Avg. Weight
Objective	Gulf	186	8.0	4.3	0.96	1.00
Objective	Non-Gulf	1484	139.6	9.4	0.91	1.03
Watches	Gulf	369	24.4	6.6	0.93	0.63
Watches	Non-Gulf	1511	111.8	7.3	0.93	0.55

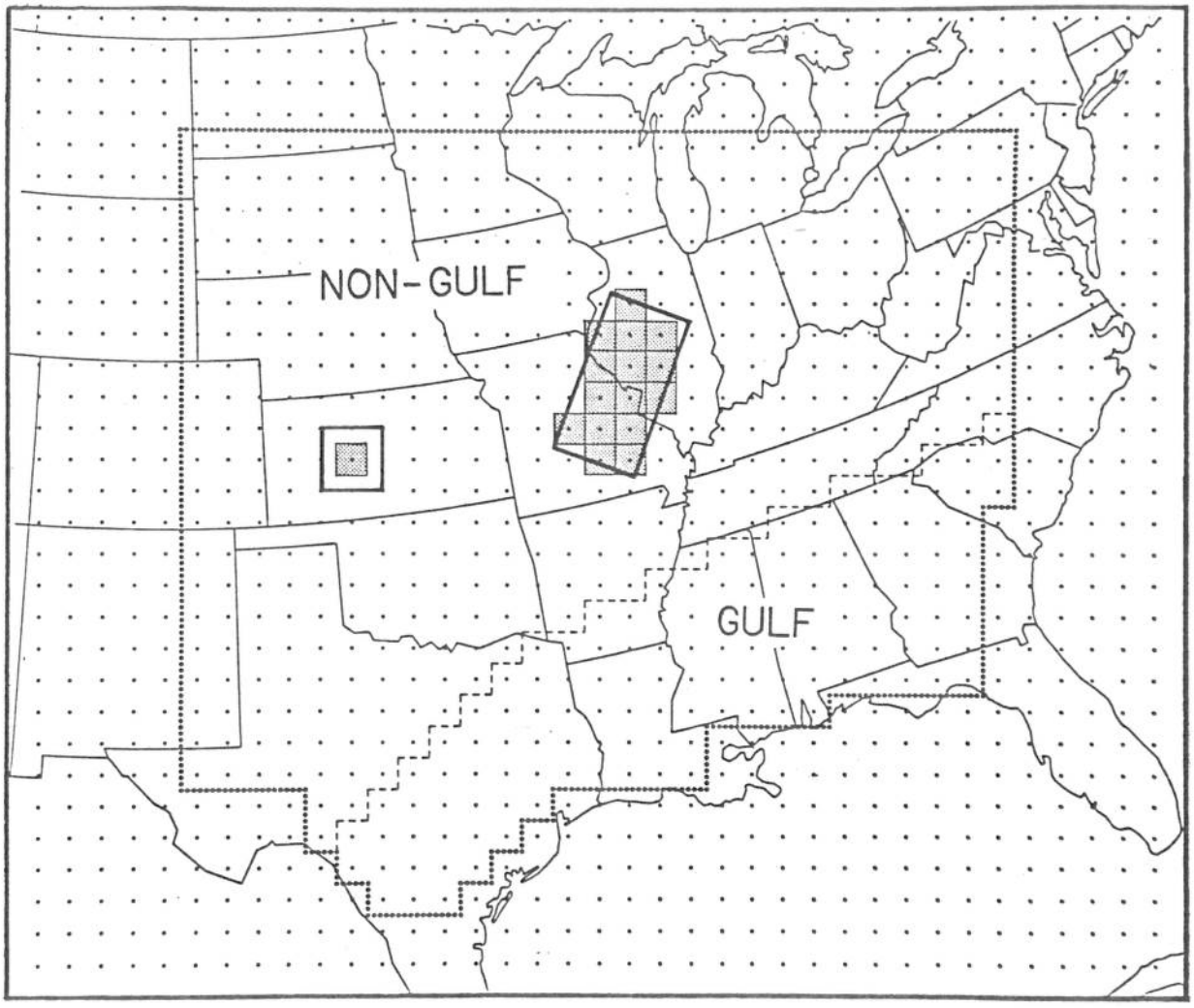

Figure 1. Areas involved in the comparative verification. The total area, enclosed by the heavy dotted line, is divided into the Gulf and non-Gulf regions as indicated by the irregular dashed line. The area for which an individual probability forecast is valid is illustrated by the larger square area. The corresponding categorical YES/NO forecast is verified over the shaded square centered inside the larger one. The example watch area shown is broken into these small square areas and each is verified as a separate YES forecast.

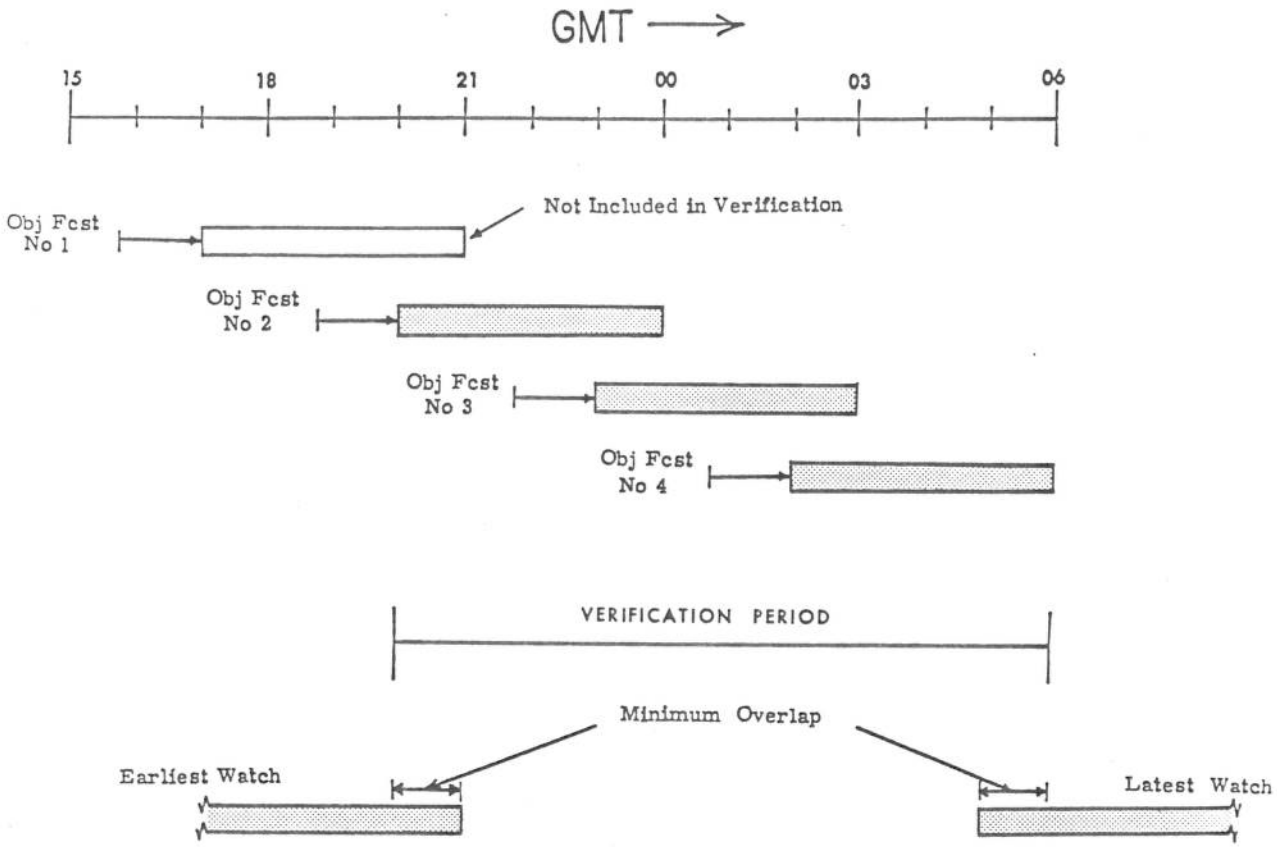

Figure 2. Valid periods of the four daily objective forecasts relative to the verification period. The issue time of an objective forecast is indicated by the short vertical line at the left end of its time schematic. The valid period of the first objective forecast of the day just fails to meet the overlap criterion and, thus, it is not included in the verification. Also shown are example watch valid periods which just satisfy the overlap criterion for inclusion.

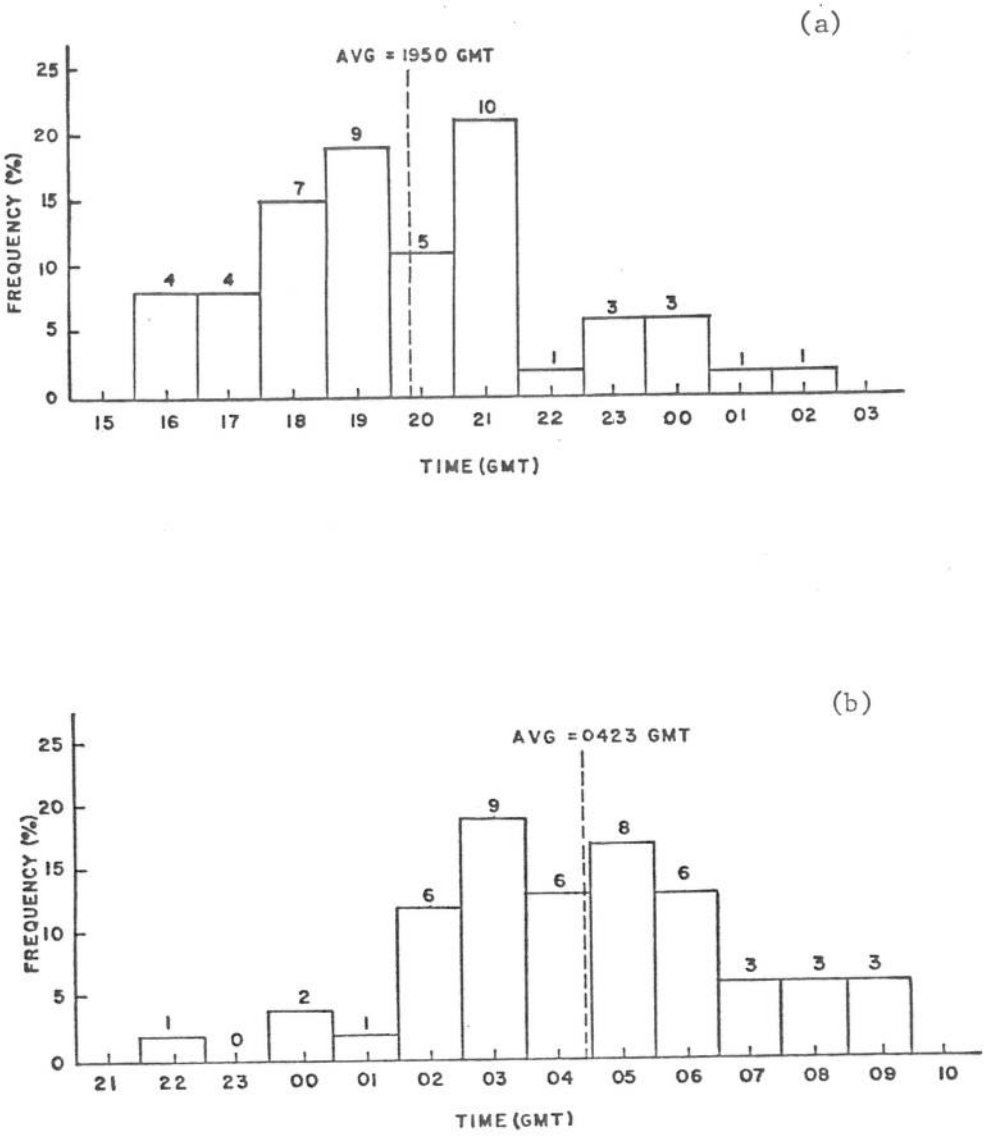

Figure 3. Frequency histograms of (a) the beginning times and (b) the ending times of the watch verification periods. The actual number of cases within each one-hour interval is given above each bar.

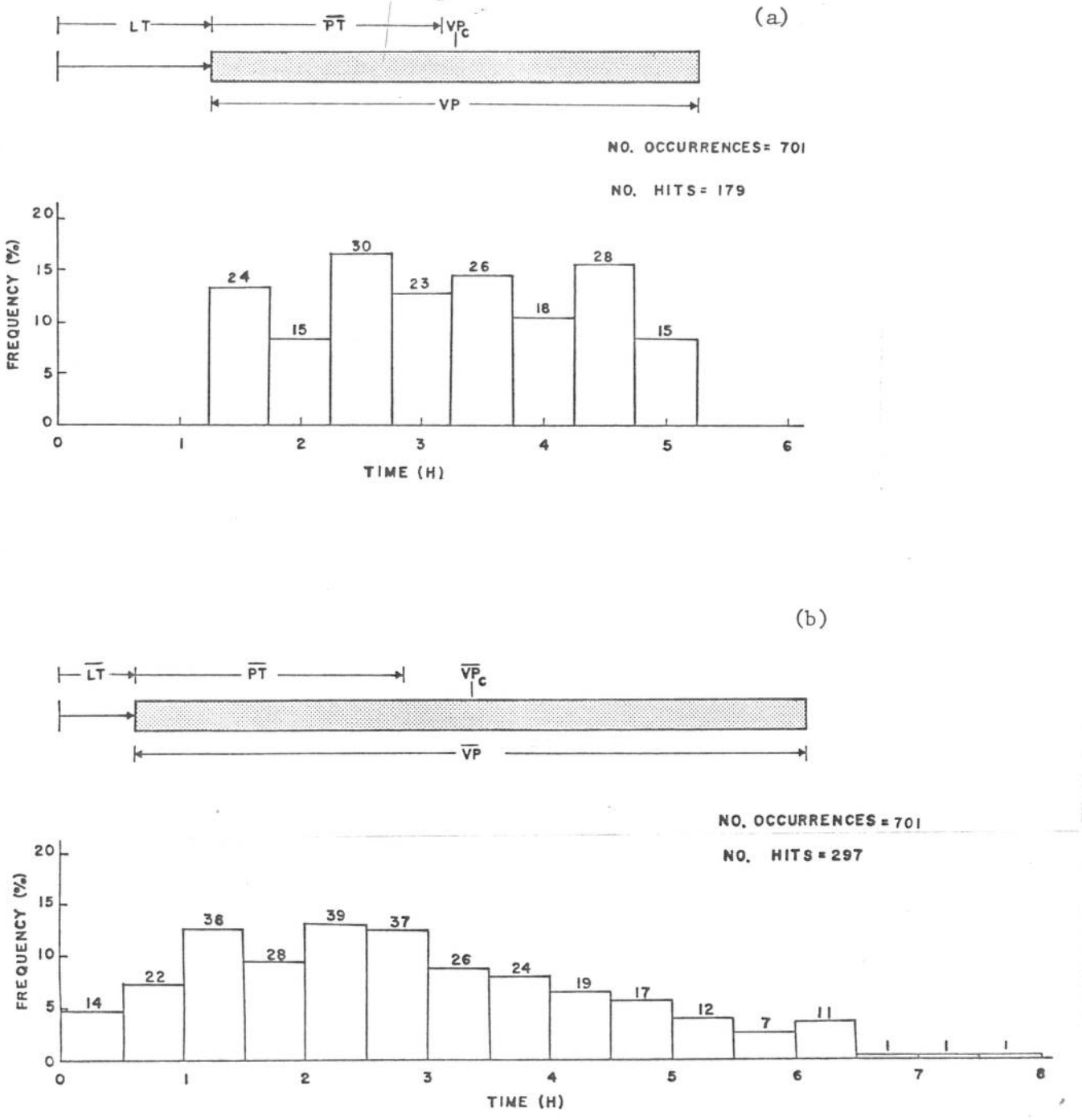

Figure 4. Frequency histograms of hits as a function of time for (a) the objective forecasts and (b) the watches in the non-Gulf region. The actual number of hits within each 30-min interval is shown. Zero along the time axis refers to the issue time of the forecast. The schematic above each histogram depicts the average lead time ($\overline{LT}$), average projection time ($\overline{PT}$), and average valid period ($\overline{VP}$). $\overline{VP}_c$ denotes the center of the average valid period. Since the LT and VP in the case of the objective forecasts is fixed, horizontal bars do not appear over these symbols in (a).

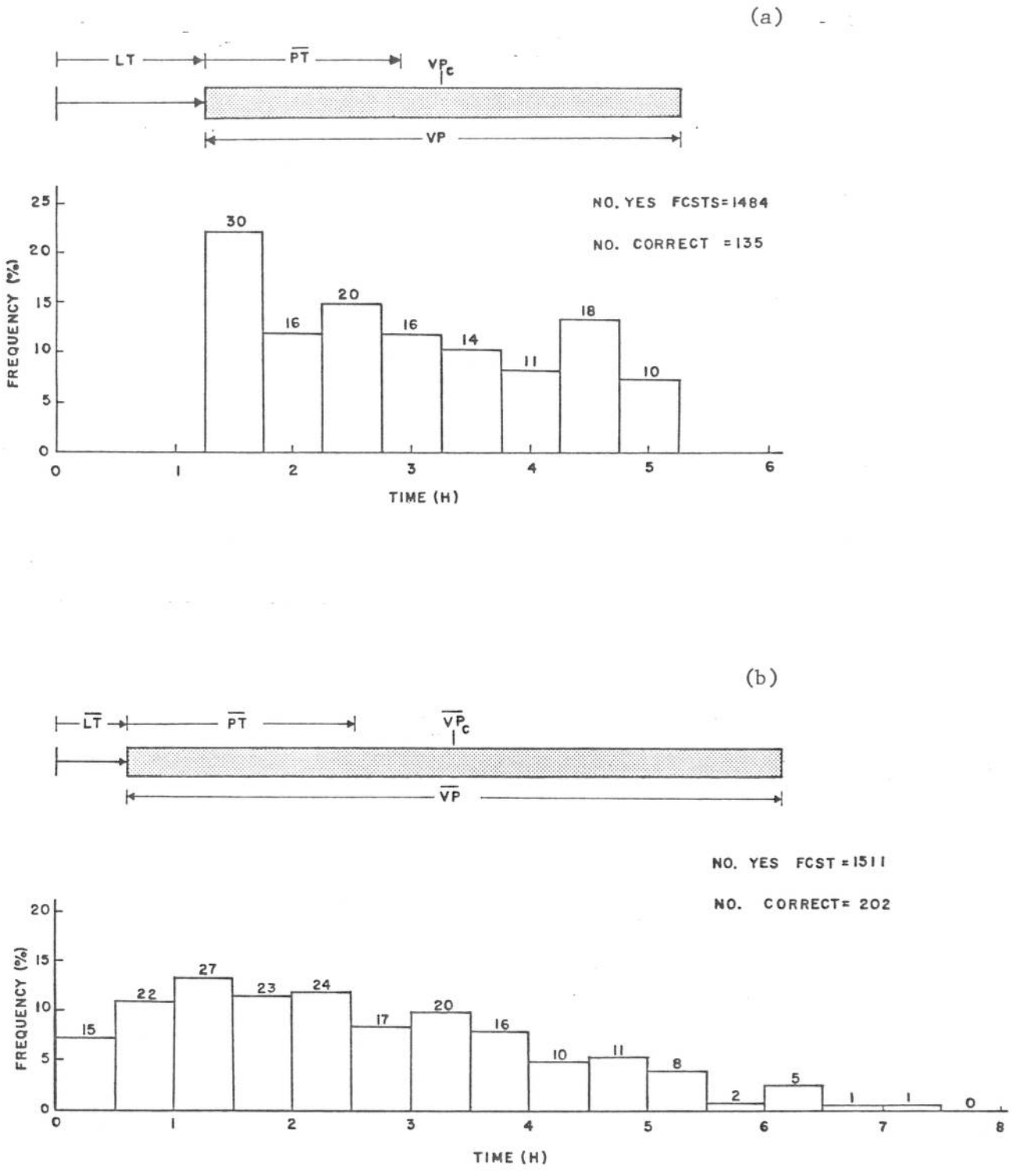

Figure 5. Same as Fig. 4 except for correct YES forecasts.